LEAF: A Benchmark for Federated Settings

Sebastian Caldas[1], Karthik Duddu[1], Peter Wu[1], Tian Li[1], Jakub Konečný[2], H. Brendan McMahan[2], Virginia Smith[1], Ameet Talwalkar[1,3]

[1]Carnegie Mellon University, [2]Google, [3]Determined AI

Find us at
leaf.cmu.edu

Why LEAF?

- Modern federated networks present new challenges in research areas related to distributed learning, meta-learning and privacy.
- We need to ensure that developments made in these areas are grounded with realistic benchmarks.
- To this end, we propose LEAF, a modular benchmarking framework for learning in federated settings.

LEAF

- LEAF includes a suite of open-source federated datasets, a rigorous evaluation framework, and a set of reference implementations (see Fig. 1).

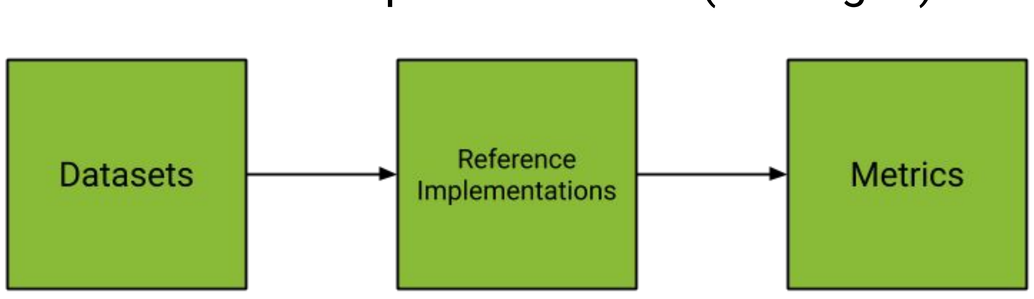

Figure 1: LEAF modules. The *Datasets* module preprocesses the data and transforms it into a standardized format. The *Reference Implementations* module is a growing repository of common methods used in the federated setting, with each implementation producing a log of various different statistical and systems metrics. Any log generated in an appropriate format can then be used to analyze these metrics in various ways through the *Metrics* module.

- All components are geared towards capturing the obstacles and intricacies of practical federated environments.

Datasets

Dataset	Number of devices	Number of samples	Task
FEMNIST	3,550	805,263	Image classification
Sent140	660,120	1,600,498	Sentiment analysis
Shakespeare	1,129	4,226,158	Next character pred.
CelebA	9,343	200,288	Image classification
Reddit	1,660,820	56,587,343	Language modeling
Synthetic	Set by user	Random	Classification

Table 1: Specifications of LEAF's *Datasets* module.

References:
[1] Brendan McMahan et al., Communication-efficient learning of deep networks from decentralized data. arXiv:1602.05629, 2016.
[2] Alex Nichol et al., On first-order meta-learning algorithms. arXiv:1803.02999, 2018.

LEAF in Action

- **LEAF enables reproducible science:** We qualitatively reproduce the results in [1], where it was noted that FedAvg diverged for the Shakespeare dataset as the number of local epochs increases (see Fig. 2).

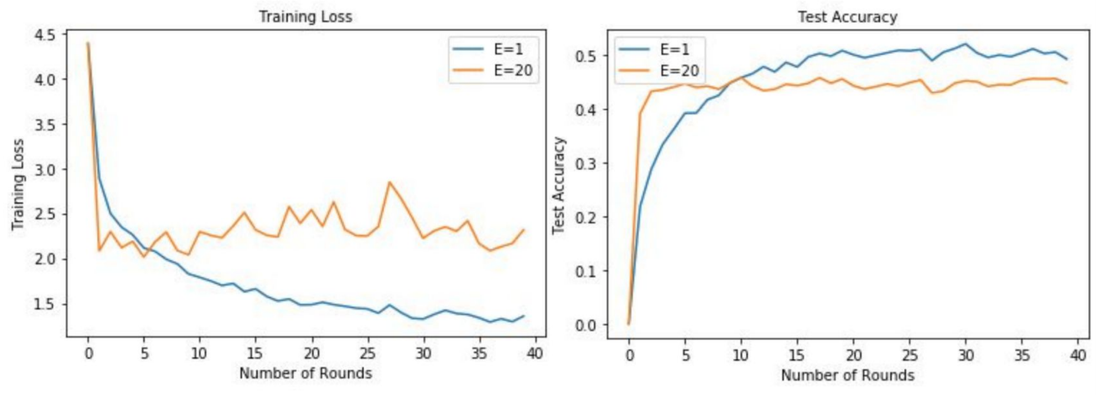

Figure 2: Convergence behavior of FedAvg on Shakespeare. Understanding this behavior is critical for the successful deployment of federated systems.

- **LEAF provides granular metrics:** Our proposed systems and statistical metrics are important to consider when serving heterogeneous clients in a distributed environment (see Fig. 3).

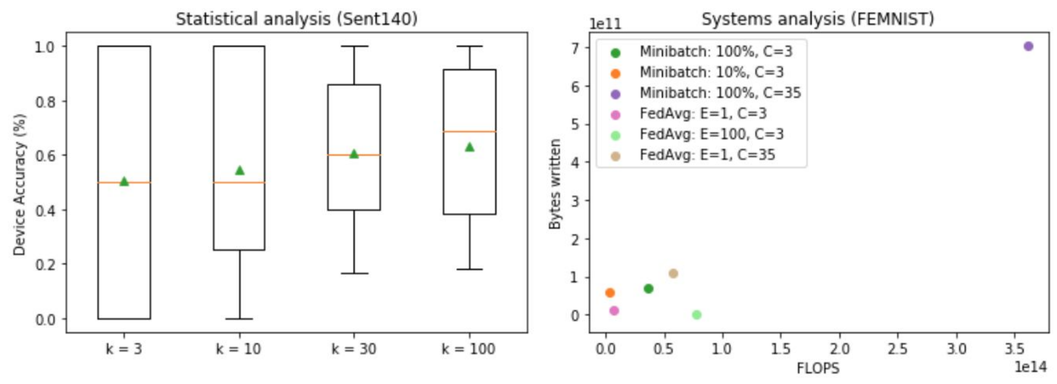

Figure 3: For statistical metrics (left), we show the effect of varying the minimum number of samples per user *k*. While median performance degrades only slightly with data-deficient users, the 25th percentile degrades dramatically. Meanwhile, for systems metrics for FEMNIST (right), our results demonstrate the improved systems profile of FedAvg over minibatch SGD when reaching a target accuracy.

- **LEAF is modular:** We can we incorporate LEAF's *Datasets* module into new experimental pipelines (see Table 2).

Dataset	FedAvg	Additional Pipeline	
		Description	Accuracy
CelebA	89.46%	Local models	65.29%
Reddit	13.35%	IID model	12.60%
FEMNIST	74.72%	Reptile [2]	80.24%

Table 2: Additional pipelines for LEAF's datasets.